Υπολογιστική Νοημοσύνη

Ιωάννης Γ. Τσούλος

Τμήμα Πληροφορικής και τηλεπικοινωνιών
Πανεπιστήμιο Ιωαννίνων

2024

Περίληψη

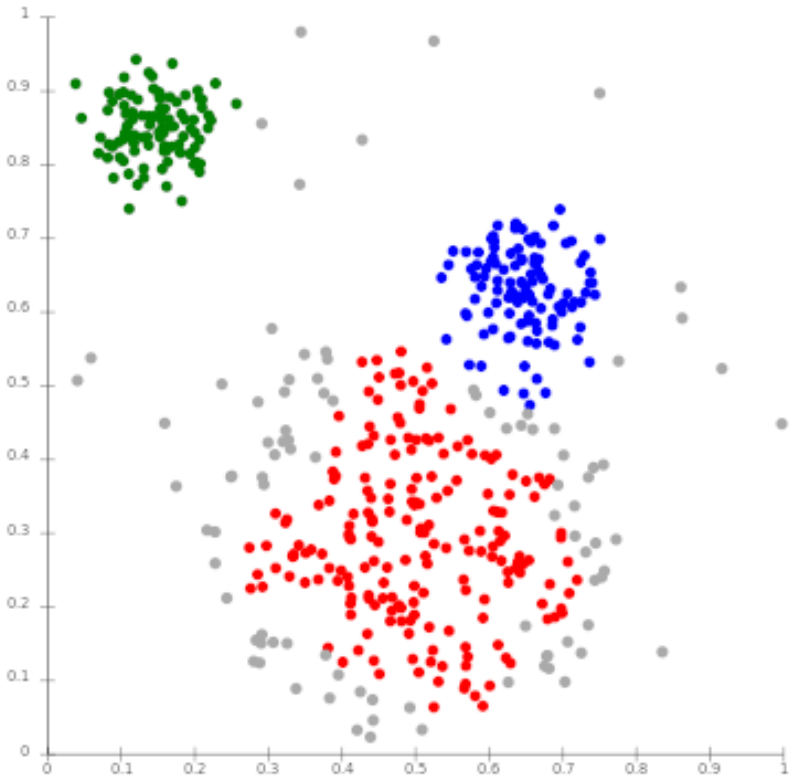
Ορισμός

Στην ομαδοποίηση "έξυπνοι" αλγόριθμοι χρησιμοποιούνται για την κατάταξη δεδομένων σε ένα προκαθορισμένο αριθμό κατηγοριών. Για παράδειγμα αν διαθέτουμε πολλά δείγματα κρασιών και θέλουμε να τα κατατάξουμε σε κατηγορίες με βάση χαρακτηριστικά τους (πχ οξύτητα, χρώμα κτλ)

Προϋποθέσεις

1. Η ομάδα να ειναι ομοιογενής, δηλαδή τα στοιχεία που απαρτίζουν μια ομάδα να είναι όσο το δυνατόν πιο κοντά μεταξύ τους
2. Οι ομάδες να απέχουν, δηλαδή να μην είναι "κοντά", γιατί αλλιώς θα πρέπει να ενωθούν σε μια.

Παράδειγμα



Παράδειγμα ομαδοποίησης

Παραδείγμα.

- Δεδομένα που ανήκουν σε τρεις κατηγορίες
- Σκοπός της μεθόδου θα πρέπει να είναι ο διαχωρισμός των γκρίζων σημείων σε κάποια από τις περιοχές αυτές
- Κάποια σημεία είναι ξεκάθαρο σε ποια ομάδα πρέπει να μπουν
- Κάποια σημεία βρίσκονται ανάμεσα στις περιοχές και δεν είναι ξεκάθαρο σε ποια περιοχή πρέπει να μπουν
- Σε κάποιες περιπτώσεις ενδεχομένως να χρειαστεί να αλλαξει το πλήθος των ομάδων (αύξηση ή μείωση).

Ορισμοί.

Για να μπορέσουμε να αξιολογήσουμε πόσο κοντά βρίσκονται τα δεδομένα θα πρέπει να υπάρχει κάποιο κριτήριο ομοιότητας. Σε όλες τις εκφράσεις που ακολουθούν ο αριθμός n εκφράζει την διάσταση (πλήθος χαρακτηριστικών) κάθε προτύπου. Μερικά γνωστά κριτήρια παρουσιάζονται στην συνέχεια.

Ευκλείδια απόσταση.

- Σε όλες τις εκφράσεις θεωρούμε πως στους τύπους έχουμε διανύσματα
- Η μεταβλητή n αναπαριστά την διάσταση του προβλήματος (διάσταση προτύπων)
- Είναι το πιο γνωστό κριτήριο απόστασης.

$$D(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{1}$$

Πίνακας ευκλίδειας απόστασης

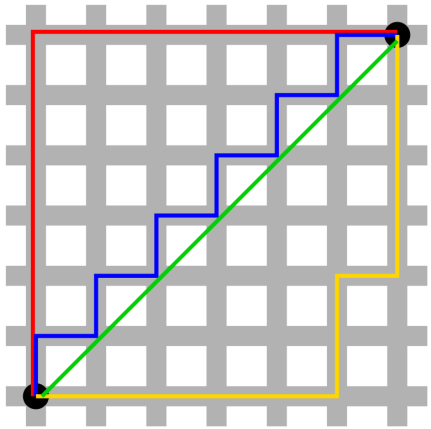
1. Είναι ο πίνακας των αποστάσεων μεταξύ διανυσμάτων.
2. Χρησιμοποιείται σε πολλές μεθόδους.
3. $A = (a_{ij}), \ a_{ij} = d_{ij}^2 = \|x_i - x_j\|^2$
4. $A_{ij} = 0, \ \forall i = j$
5. $A_{ij} = A_{ji}$, συμμετρικός πίνακας
6. $A_{ij} \geq 0$

Απόσταση Manhattan.

1. Βασίζεται σε παρατηρήσεις σχετικά με τις αποστάσεις στην τετραγωνισμένη περιοχή του Μανχάττσαν
2. Η απόσταση αυτή δίνεται από την εξίσωση

$$D(x, y) = \sum_{i=1}^{n} |x_i - y_i| \tag{2}$$

Απόσταση Manhattan (σχήμα)



Κόκκινο χρώμα: Απόσταση Manhattan distance. Πράσινο: Απευθείας μετακίνηση. Λοιπά χρώματα: Ισοδύναμες αποστάσεις Manhattan

Μέγιστης διαφοράς.

1. Το κριτήριο αυτό βασίζεται στην εύρεση της μέγιστης διαφοράς σε όλες τις διαστάσεις των προτύπων
2. Χρησιμοποιείται αρκετά τακτικά όπως και της Ευκλείδιας απόστασης.
3. Δίνεται από την εξίσωση

$$D(x, y) = max_{i=1}^{n} |x_i - y_i| \tag{3}$$

Συνημιτονοειδής ομοιότητα.

Ορίζεται από την εξίσωση

$$D(x,y) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \sqrt{\sum_{i=1}^{n} y_i^2}} \tag{4}$$

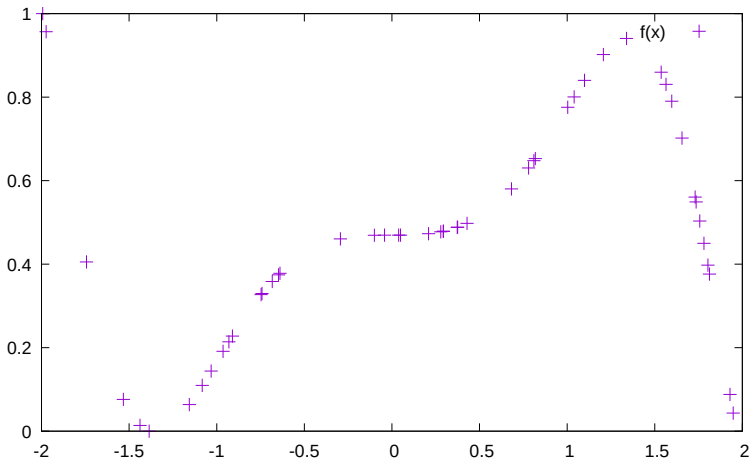
Ορισμοί

- Η μέθοδος αναπτύχθηκε από τους Fix και Hodges το 1951[1].
- Μη παραμετρική μέθοδος, δηλαδή δεν υπάρχουν παράμετροι που πρέπει να εκτιμηθούν.
- Τα δεδομένα κατατάσσονται στην κατηγορία που πλειοψηφεί ανάμεσα στους K κοντινότερους γείτονες τους.
- Παίζει σημαντικό ρόλο η τιμή του K (αριθμός γειτόνων)
- Παίζει λιγότερο σημαντικό ρόλο το είδος της απόστασης που θα χρησιμοποιηθεί.
- Μπορεί να χρησιμοποιηθεί τόσο για κατηγοριοποίηση δεδομένων όσο και για μάθηση συναρτήσεων.
- Είναι ανεκτικός αλγόριθμος σε παρουσία θορύβου και όταν ακόμα λείπουν τιμές από χαρακτηριστικά.

KNN για κατηγοριοποίηση

1. Δημιουργία συνόλου εκπαίδευσης $S = \{X_1, X_2, ..., X_N\}$, όπου $X_i \in R^d$
2. Καθορισμός της παραμέτρου K. Συνήθως οι τιμές αυτής της παραμέτρου είναι μονοί αριθμοί.
3. Για κάθε νέο πρότυπο X_i
   1. Δημιουργία του συνόλου S_x με τους K κοντινότερους γείτονες από το σύνολο S. Για την εύρεση των γειτόνων χρησιμοποιούνται διάφορα κριτήρια απόστασης με το πιο συνηθισμένο την **Ευκλείδια** απόσταση.
   2. Εύρεση της κατηγορίας Υ που πλειοψηφεί στο σύνολο S_x
   3. Ανάθεση του προτύπου στην κατηγορία Υ.
4. Ο αλγόριθμος βασίζεται στο K. Επίσης είναι σχετικά αργός αλγόριθμος, αφού απαιτεί ταξινόμηση για κάθε πρότυπο.
5. Πιθανή λύση η δημιουργία πίνακα απόστάσεων (όπως ο πίνακας Ευκλείδιας απόστασης).

Μάθηση συναρτήσεων

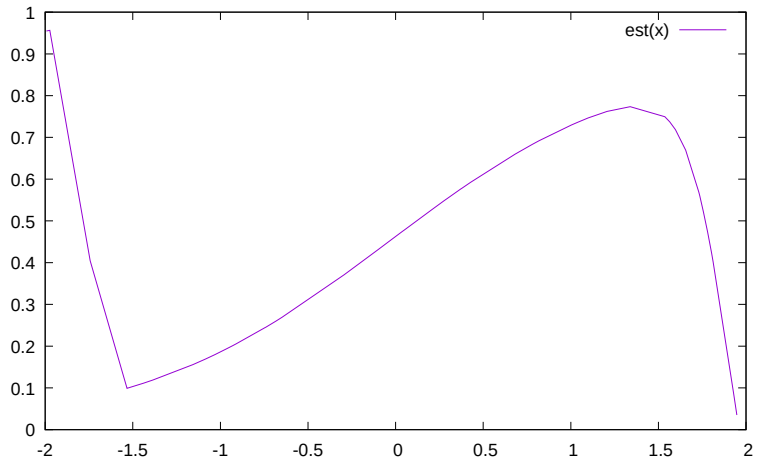
1. Με τον όρο μάθηση συναρτήσων μιλάμε για εύρεση της καμπύλης (συνάρτησης) που πιθανόν να βρίσκεται πίσω από δεδομένα.
2. Σκοπός ενός μοντέλου που κάνει μάθηση συναρτήσεων είναι η εκτίμηση της συνάρτησης που περνά από αυτά τα σημεία αλλά και από άλλα ενδιάμεσα σημεία και πιθανόν και από σημεία εκτός του διαστήματος.
3. 

Χρήση KNN για μάθηση συναρτήσεων

1. Δημιουργία συνόλου εκπαίδευσης $S = \{X_1, X_2, ..., X_N\}$, όπου $X_i \in R^d$
2. Καθορισμός της παραμέτρου K. Συνήθως οι τιμές αυτής της παραμέτρου είναι μονοί αριθμοί.
3. Για κάθε νέο πρότυπο X_i
   1. Δημιουργία του συνόλου S_x με τους K κοντινότερους γείτονες από το σύνολο S. Για την εύρεση των γειτόνων χρησιμοποιούνται διάφορα κριτήρια απόστασης με το πιο συνηθισμένο την **Ευκλείδια** απόσταση.
   2. Υπολογισμός της τιμής

$$Y(x) = \frac{1}{K} \sum_{i=1}^{K} X_i, \forall\ X_i \in S_x$$

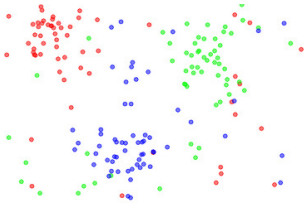
Παράδειγμα μάθησης συναρτήσεων



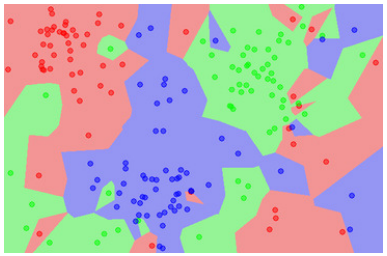
Για το προηγούμενο σύνολο δεδομένων με χρήση KNN

Παράδειγμα εύρεσης κατηγοριών

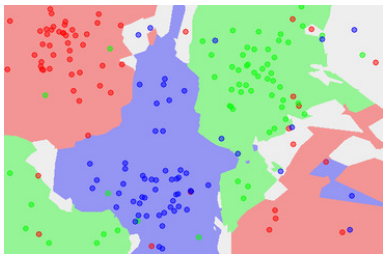
1. Αρχικό σύνολο δεδομένων



2. Χρήση ενός γείτονα



3. Χρήση 5 γειτόνων



Ιδιότητες

1. Για K=1 δεν έχουμε τόσο καλή συμπεριφορά, καθώς κάνει πολλά λάθη
2. Για μεγαλύτερες τιμές του K, με K περιττό παρατηρείται καλύτερη συμπεριφορά
3. Για πολύ μεγάλες τιμές γίνονται πάλι λάθη, καθώς συμμετέχουν στην ψηφοφορία και πολύ μακρινά σημεία

KNN με χρήση βαρών

- Έχει παρουσιάσει καλύτερες ικανότητες μάθησης σε αρκετά παραδείγματα
- Συσχετίζουμε κάθε γείτονα με ένα βάρος
- Έχουν αναπτυχθεί αρκετές τεχνικές KNN με βάρη
- Στην συνέχεια παρουσιάζεται η τεχνική Inverse Weighted KNN[2]

KNN με βάρη

1. Για ένα πρότυπο x_i εύρεση των αποστάσεων $d_j, j = 1..K$
2. Για κάθε απόσταση d_j υπολογισμός της ποσότητας $V_j = \frac{1}{d_j}$
3. Υπολογισμός των βαρών $w_j = \frac{V_j}{\sum V_k}$
4. Ανάθεση του προτύπου στην κατηγορία με το μεγαλύτερο άθροισμα βαρών

Ομαδοποίηση πλησιέστερων γειτόνων

1. Για κάθε πρότυπο x_i δημιούργησε την λίστα $L(x_i)$ με τους k κοντινότερους γείτονες.
2. Για κάθε ζεύγος σημείων x_i και x_j
   1. Αν $L(x_i) \cap L(x_j) \geq M$, τοποθέτησε τα δύο σημεία x_i και x_j στην ίδια ομάδα
3. Η διαδικασία επαναλαμβάνεται μέχρι να μην υπάρχουν πλέον άλλα σημεία εκτός ομάδας.

Αυτός ο αλγόριθμος στηρίζεται στις παραμέτρους k και M

Regression Error

1. Είναι το σφάλμα που μας ενδιαφέρει για μάθηση συναρτήσεων (προσαρμογή δεδομένων)
2. Χρησιμοποιείται για να μετρήσει το σφάλμα ενός μοντέλου μάθησης $\Psi(x)$
3. Σκοπός του μοντέλου είνα να "μαντέψει" τις σωστές εξόδους $y_i,\ i = 1, .., m$
4. Ορίζεται ως

$$E_R(\Psi(x)) = \frac{1}{m}\sum_{i=1}^{m}\left(\Psi\left(x_i\right) - y_i\right)^2$$

5. Αποτυπώνει το μέσο τετραγωνικό σφάλμα ανά σημείο.

Classification Error

1. Μετράει την μέση διαφοροποίηση σε προβλήματα ταξινόμησης ανάμεσα στην εκτιμούμενη κατηγορία και την πραγματική
2. Θεωρούμε το μοντέλο $\Psi(x)$
3. Το μοντέλο είτε παράγει κατηγορία είτε την εκτιμούμε με χρήση κατωφλίων πχ $C(x)$
4. Ορίζεται ως

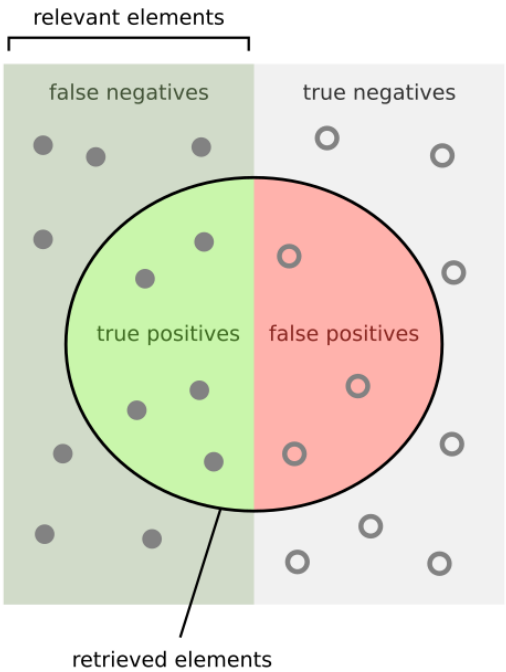
$$E_C(\Psi(x)) = \frac{1}{m}\sum_{i=1}^{n}(C(\Psi(x_i)) \neq y_i)$$

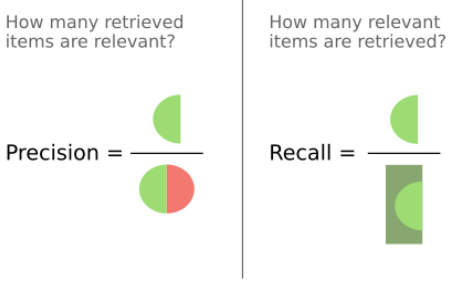
Confusion matrix

1. Αν έχουμε N κατηγορίες τότε έχει διάσταση $N \times N$
2. C_{ij}=Πλήθος προτύπων που ενώ ανήκουν στην κατηγορία i ανατέθηκαν στην κατηγορία j
3. Άθροισμα στοιχείων κύριας διαγωνίου: Πόσα στοιχεία έχουν ταξινομηθεί σωστά
4. Άθροισμα των υπολοίπων στοιχείων: Πόσα στοιχεία έχουν ταξινομηθεί λάθος

Precision και Recall

1. Precision: $p = \frac{tp}{tp+fp}$
2. Recall: $r = \frac{tp}{tp+fn}$
3. Σχηματικά (2 κατηγορίες):





Ορισμοί

1. Δημιουργία εκπροσώπων από ομάδες.
2. Έχει υλοποιηθεί αρχικά από τον MacQueen[3].
3. Το πλήθος των ομάδων θεωρείται δεδομένο.
4. Έχουν αναπτυχθεί δεκάδες παραλαγές του αλγορίθμου από τότε.

Ο αλγόριθμος

1. **Αρχικοποίηση** των K κέντρων $c_i,\ i = 1..K$, όπου K είναι το εκτιμούμενο πλήθος ομάδων. Κάθε κέντρο c_i θεωρούμε πως έχει n στοιχεία, όπου n είναι η διάσταση των προτύπων εισόδου.
2. **Επανέλαβε**
   1. $S_i = \{\},\ i = 1..K$
   2. Εύρεση της ομάδας που ανήκει το κάθε στοιχείο $x_i,\ \iota = 1..\mathrm{N}$: α) εύρεση $j^* = \min_{i=1}^{K} \{D(x_i, c_j)\}$ β) $S_{j^*} = S_{j^*} \cup x_i$
   3. Ανανέωση του κέντρου της ομάδας

$$c_j = \frac{1}{M_j} \sum_{x_i \in S_j} x_i \tag{5}$$

   όπου M_j το πλήθος των μελών της ομάδας j.
3. **Αν** τα κέντρα δεν έχουν αλλάξει **τότε τερματισμός**, **αλλιώς** μετάβαση στο βήμα 2.

Σύνοψη

- Παρουσιάστηκαν η έννοια της κατηγοριοποίησης
- Παρουσιάστηκε ο αλγόριθμος KNN και οι επεκτάσεις του
- Δόθηκε μια σύντομη παρουσίαση του αλγορίθμου KMEANS

Βιβλιογραφία I


Fix, Evelyn; Hodges, Joseph L. (1951). Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties (PDF) (Report). USAF School of Aviation Medicine, Randolph Field, Texas.


https://visualstudiomagazine.com/articles/2019/04/01/weighted-k-nn-classification.aspx


MacQueen, J.: Some methods for classification and analysis of multivariate observations, in: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Vol. 1, No. 14, pp. 281-297, 1967.